Descriptive Statistics

The **DESCRIPTIVE STATISTICS** procedure displays univariate summary statistics for selected variables. Descriptive statistics can be used to describe the basic features of the data in a study. It provides simple summaries about the sample and the measures. Together with simple graphical analysis, it can form the basis of quantitative data analysis.

How To

- ✓ Run **STATISTICS->BASIC STATISTICS->DESCRIPTIVE STATISTICS.**
- ✓ Select one or more variables.
- ✓ Optionally, use the **PLOT HISTOGRAM** option to build a histogram with frequencies and normal curve overlay for each variable. Normal curve overlay is not available when a report is viewed in Apple Numbers, because of a lack of combined charts support in the Apple Numbers app.
- ✓ By default, a table with descriptive statistics is produced for each variable. To view descriptive statistics for all variables in a single table – select the "Single table" value for the Report option.

Report: For each variable

Variable #1 (Var1-x)		
Count	11.00	Mean Devi
Mean	9.	Second Mo
Mean LCL	6.23623	Third Mon
Mean UCL	11.76377	Fourth Mo
Variance	11.	
Standard Deviation	3.31662	Sum
Mean Standard Error	1.	Sum Stand
Coefficient of Variation	0.36851	Total Sum
		Adjusted S
Minimum	4.	
Maximum	14.	Geometric
Range	10.	Harmonic
		Mode
Median	9.	
Median Error	0.37789	Skewness
Percentile 25% (Q1)	6.5	Skewness
Percentile 75% (Q3)	11.5	Kurtosis
IQR	5.	Kurtosis St
MAD (Median Absolute Deviation)	5.	Skewness
Coefficient of Dispersion (COD)	0.30303	Kurtosis (F
Variable #2 (Var1-y)		
Count	11.00	Mean Devi
Mean	7.50091	Second Mo
Mean LCL	5.80799	Third Mon
Mean UCL	9.19383	Fourth Mo
Variance	4.12727	
Standard Deviation	2.03157	Sum
Mean Standard Error	0.61254	Sum Stand
Coefficient of Variation	0.27084	Total Sum
		Adjusted S

Report: Single table

	Var1-x	Var1-y
Count	11.00	11.00
Mean	9.	7.50091
Mean LCL	6.23623	5.80799
Mean UCL	11.76377	9.19383
Variance	11.	4.12727
Standard Deviation	3.31662	2.03157
Mean Standard Error	1.	0.61254
Coefficient of Variation	0.36851	0.27084
Minimum	4.	4.26
Maximum	14.	10.84
Range	10.	6.58
Median	9.	7.58
Median Error	0.37789	0.23147
Percentile 25% (Q1)	6.5	6.315
Percentile 75% (Q3)	11.5	8.57
IQR	5.	2.255
MAD (Median Absolute Deviation)	5.	2.38
Coefficient of Dispersion (COD)	0.30303	0.20425

In single table view the first column is frozen, so you can scroll through the report while the heading column stays still.

- ✓ Optionally, select a method for computing percentiles. Percentiles are defined according to Hyndman and Fan (1996), see below for details.

Results

Table with summary statistics is produced for each variable. The table includes following statistics.

COUNT (N) - sample size.

MEAN – arithmetic mean. The larger the sample size, the more reliable is its mean. The larger the variation of data values, the less reliable the mean.

$$\bar{x} = \frac{1}{N}\sum_{1}^{N} x_i$$

MEAN LCL, MEAN UCL – are the lower value (LCL) and upper value (UCL) of $(1-\alpha)\%$ reliable interval limits estimate for the mean based on a t-distribution with $N-1$ degrees of freedom. The estimates are made assuming that the population standard deviation is not known and that the variable is normally distributed.

$$Lower\ limit\ =\ \bar{x} -\ t_{CL}\ Sm$$
$$Upper\ limit\ =\ \bar{x} +\ t_{CL}\ Sm$$

t_{CL} – t for the $(1-\alpha)\%$ confidence level (default value = 95%, default $\alpha = 0.05$). α can be changed in the **PREFERENCES**.
Sm – estimated standard error of the mean.

LCL is for **L**ower **C**onfidence **L**imit and UCL is for **U**pper **C**onfidence **L**imit.

VARIANCE (UNBIASED ESTIMATE) - is the mean value of the square of the deviation of that variable from its mean with Bessel's correction.

$$s^2 = \frac{1}{N-1}\sum_{1}^{N}(x_i - \bar{x})^2$$

Population variance is estimated as

$$\sigma^2 = \frac{1}{N}\sum_{1}^{N}(x_i - \bar{x})^2 = \mu_2,$$

where μ_2 is second moment (see below).

STANDARD DEVIATION - square root of the variance.

$$\sigma = \sqrt{\frac{1}{N}\sum_{1}^{N}(x_i - \bar{x})^2}$$

STANDARD ERROR (OF MEAN) - quantifies the precision of the mean. It is a measure of how far your sample mean is likely to be from the true population mean. The formula shows that the larger the sample size, the smaller the standard error of the mean. More specifically, the size of the standard error of the mean is inversely proportional to the square root of the sample size.

$$SEM = \frac{\sigma}{\sqrt{N}}$$

MINIMUM – the smallest value for a variable.

MAXIMUM – the largest value for a variable.

RANGE - difference between the largest and smallest values of a variable. For normally distributed variable dividing the range by six can make a quick estimate of the standard deviation.

SUM – sum of the sample values.

SUM STANDARD ERROR - standard deviation of sums distribution.

TOTAL SUM SQUARES - the sum of the squared values of the variable. Sometimes referred to as *the unadjusted sum of squares*.

$$TSS = \sum_{1}^{N} {x_i}^2$$

ADJUSTED SUM SQUARES - the sum of the squared differences from the mean.

$$AdjSS = \sum_{1}^{N} (x_i - \bar{x})^2$$

GEOMETRIC MEAN - a type of mean, which indicates the central tendency of a set of numbers. It is similar to the arithmetic mean, except that instead of adding observations and then dividing the sum by the count of observations *N*, the observations are multiplied, and then the n^{th} root of the resulting product is taken. Geometric mean is used to find average rates of change, average rates of growth or average ratios.

$$G = \sqrt[N]{\prod_{1}^{N} x_i}$$

HARMONIC MEAN - or subcontrary mean, the number H defined as

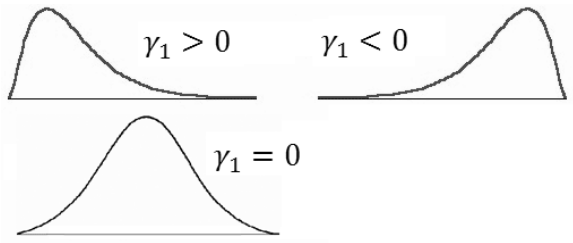
$$H = \frac{1}{N}\sum_{1}^{N}\frac{1}{x_i}.$$

As seen from the formula above, harmonic mean is the reciprocal of the arithmetic mean of the reciprocals. Harmonic mean is used to calculate an average value when data are measured as a rate, such as ratios (price-to-earnings ratio or P/E Ratio), consumption (miles-per-gallon or MPG) or productivity (output to man-hours).

MODE - the value that occurs most frequently in the sample. The mode is a measure of central tendency. It is not necessarily unique since the same maximum frequency may be attained at different values (in this case #N/A is displayed).

SKEWNESS – a measure of the asymmetry of the variable. A value of zero indicates a symmetrical distribution, i.e. **Mean** = **Median**. The typical definition is:

$$\gamma_1 = \sum_{1}^{N}\frac{\mu_3}{\sigma^{3/2}} = \frac{1}{N}\sum_{1}^{N}(\frac{x_i - \bar{x}}{\sigma})^3$$



There are different formulas for estimating skewness and kurtosis (Joanes, Gill, 1998). The formula above is used in many textbooks and some software packages (NCSS, Wolfram Mathematica). Use the **SKEWNESS (FISHER'S)** *value to get the same results as in SPSS, SAS and Excel software.*

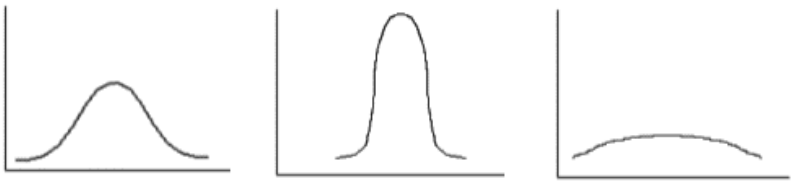
SKEWNESS STANDARD ERROR – large sample estimate of the standard error of skewness for an infinite population.

$$k_1 = \frac{\gamma_1}{\sigma^3}$$

KURTOSIS - a measure of the "peakedness" of the variable. Higher kurtosis means more of the variance is the result of infrequent extreme deviations, as opposed to frequent modestly sized deviations. If the kurtosis equals three and the skewness is zero, the distribution is normal.

$$\gamma_2 = \sum_{1}^{N}\frac{\mu_4}{\sigma^2} = \frac{1}{N}\sum_{1}^{N}\left(\frac{x_i - \bar{x}}{\sigma}\right)^4$$

If $\gamma_2 = 3$ – the distribution is mesokurtic.
If $\gamma_2 > 3$ – the distribution is leptokurtic.
If $\gamma_2 < 3$ – the distribution is platykurtic.



$\gamma_2 = 3$ $\gamma_2 > 3$ $\gamma_2 < 3$

Biased estimate for kurtosis is

$$\gamma_2 = \sum_1^N \frac{\mu_4}{\sigma^2} - 3 = \frac{1}{N}\sum_1^N \left(\frac{x_i - \bar{x}}{\sigma}\right)^4 - 3$$

There are different formulas for estimating skewness and kurtosis (Joanes, Gill, 1998). The formula above is used in many textbooks and some software packages (NCSS, Wolfram Mathematica). Use the ***KURTOSIS (FISHER'S)*** *value to get the same results with SPSS, SAS and Excel software.*

KURTOSIS STANDARD ERROR - large sample estimate of the standard error of kurtosis for an infinite population.

$$k_2 = 2k_1\sqrt{\frac{n^2-1}{(n-3)(n+5)}}$$

SKEWNESS (FISHER'S) – a bias-corrected measure of skewness. Also known as **FISHER'S SKEWNESS G1.**

$$g_1 = \frac{\sqrt{n(n-1)}}{n-2}\gamma_1$$

KURTOSIS (FISHER'S)- an alternative measure of kurtosis based on the unbiased estimators of moments. Also known as **FISHER'S KURTOSIS G2.**

$$g_2 = \frac{(n+1)(n-1)}{(n-2)(n-3)}\left\{\gamma_2 - 3\frac{n-1}{n+1}\right\}$$

COEFFICIENT OF VARIATION - a normalized measure of dispersion of a probability distribution. Defined only for non-zero mean, and is most useful for variables that are always positive. It is also known as *unitized risk* or the *variation coefficient*.

$$cv = \frac{\sigma}{\bar{x}}$$

MEAN DEVIATION (MEAN ABSOLUTE DEVIATION, MD) - mean of the absolute deviations of a set of data about the data's mean.

$$MD = \frac{1}{N}\sum_{1}^{N} |x_i - \bar{x}|$$

SECOND MOMENT, THIRD MOMENT, FOURTH MOMENT – central moments about the mean. A j^{th} central moment about the mean is defined as

$$\mu_j = \frac{1}{N}\sum_{1}^{N} (x_i - \bar{x})^j.$$

Second moment μ_2 is a biased variance estimate.

MEDIAN - the observation that splits the variable into two halves. The median of a sample can be found by arranging all the sample values from lowest value to highest value and picking the middle one. Unlike the arithmetic mean, the median is robust against outliers.

MEDIAN ERROR - the number defined by

$$SEM = s\sqrt{\frac{\pi}{2N}}$$

PERCENTILE 25% (Q1) - value of a variable below which 25% percent of observations fall.
PERCENTILE 75% (Q2) - value of a variable below which 75% percent of observations fall.

PERCENTILE DEFINITION

You can change the percentile calculation method in the **ADVANCED OPTIONS**. Nine methods from Hyndman and Fan (1996) are implemented. Sample quantiles are based on one or two order statistics and can be written as $Q(p) = (1-\gamma)\, X_{(j)} + \gamma\, X_{(j+1)}$, where $X_{(j)}$ is the sample order statistics and $\gamma = \gamma(j, g)$ $(0 \le \gamma \le 1)$ is a real-valued function of $j = \lfloor pN + m \rfloor$ (largest integer not greater than $pn + m$) and $g = \mathrm{frac}(pn + m)$, m – real constant.

Discontinuous definitions	
1. Inverse of EDF (SAS-3)	The oldest and most studied definition that uses the inverse of the empirical distribution function (EDF). $\begin{cases}\gamma = 1 \; if \; g > 0 \\ \gamma = 0 \; if \; g = 0\end{cases}$, $g = N \cdot p$ $(m = 0)$

2. EDF with averaging (SAS-5)	Similar to the previous definition, but averaging is used when $g = 0$. $\begin{cases}\gamma = 1 & if\ g > 0 \\ \gamma = 1/2 & if\ g = 0\end{cases}$, $g = N \cdot p\ (m = 0)$
3. Observation closest to N*p (SAS-2)	Defined as the order statistic $X_{(k)}$, where *k* is the nearest integer to $N \cdot p$.
Continuous definitions	
4. Interpolation of EDF (SAS-1)	Defined as the linear interpolation of function from the first definition, $p_k = k/N$.
5. Piecewise linear interpolation of EDF (midway values as knots)	Piecewise linear function with knots defined as values midway through the steps of the EDF, $p_k = (k - 0.5)/N$.
6. Interpolation of the expectations for the order statistics (SPSS, NIST)	Knots are defined as the order statistics expectations. In definitions 6 – 8, $F[X_{(k)}]$ has the distribution of the k[th] order statistics from a uniform distribution, namely the $\beta(k, N - k + 1)$. This definition is used by Minitab* and SPSS* packages. $p_k = \mathrm{E}\, F[X_{(k)}] = k/(N + 1)$.
7. Interpolation of the modes for the order statistics (Excel)	Linear interpolation of the order statistics modes. $p_k = \mathrm{mode}\, F[X_{(k)}] = (k - 1)/(N - 1)$.
8. Interpolation of the approximate medians for order statistics	Linear interpolation of the order statistics medians. Median position $\mathrm{M}\, F[X_{(k)}]$ is approximated as $\mathrm{M}\, F[X_{(k)}] \approx (k - {}^1/_3)/(N + {}^1/_3)$. Recommended by Hyndman and Fan (1996). $F[X_{(k)}]$ is defined the same way as in (6). $p_k = (k - {}^1/_3)/(N + {}^1/_3)$.
9. Blom's unbiased approximation	This definition, proposed by Blom (1958), is an approximately unbiased approximation of $Q(p)$, when F is normal. $p_k = (k - {}^3/_8)/(N + {}^1/_4)$.

IQR (INTERQUARTILE RANGE, MIDSPREAD) – the difference between the third quartile and the first quartile (between the 75th percentile and the 25th percentile). IQR represents the range of the middle 50 percent of the distribution. It is a very robust (not affected by outliers) measure of dispersion. The IQR is used to build box plots.

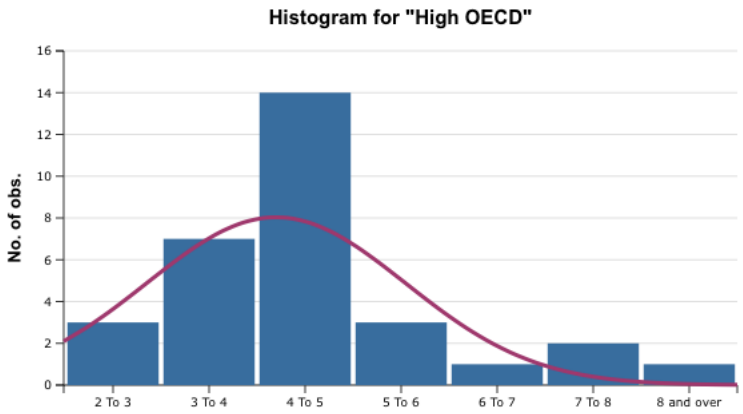
$IQR = Q3 - Q1$

MAD (MEDIAN ABSOLUTE DEVIATION) - a robust measure of the variability of a univariate sample of quantitative data. The median absolute deviation is a measure of statistical dispersion. It is a more robust estimator of scale than the sample variance or standard deviation.
$MAD = median_i\{|x_i - median_j(x_j)|\}$

COEFFICIENT OF DISPERSION – a measure of relative inequality (or relative variation) of the data. Coefficient of dispersion is the ratio of the Average Absolute Deviation from the Median (MAAD) to the Median of the data.

$$CD = \frac{1}{N}\left|\frac{MAD}{Median}\right|$$

Histogram for each variable is plotted if the corresponding option is selected in the **ADVANCED OPTIONS**. To specify the bins manually – please use the **STATISTICS->BASIC STATISTICS ->HISTOGRAM** command.



References

Blom G. (1958). Statistical estimates and transformed beta-variables. New York: Wiley.

Hyndman, R.J., Fan, Y. (November 1996). "Sample Quantiles in Statistical Packages", The American Statistician 50 (4): pp. 361–365.

Joanes, D. N., Gill, C. A. (1998), Comparing measures of sample skewness and kurtosis. The Statistician, 47, 183–189.